Statistical Analysis of Discrete-valued Time Series by Parsimonious High-order Markov Chains

Yuriy Kharin

Research Institute for Applied Problems of Mathematics and Informatics,
Belarusian State University

Abstract

Problems of statistical analysis of discrete-valued time series are considered. Two approaches for construction of parsimonious (small-parametric) models for observed discrete data are proposed based on high-order Markov chains. Consistent statistical estimators for parameters of the developed models and some known models, and also statistical tests on the values of parameters are constructed. Probabilistic properties of the constructed statistical inferences are given. The developed theory is also applied for statistical analysis of spatio-temporal data. Theoretical results are illustrated by computer experiments on real statistical data.

Keywords: discrete time series, high-order Markov chain, parsimonious model, statistical inferences.

1. Introduction

Time series analysis is deep developed (Anderson 1971) for "continuous" data when the observation space $\mathbf{A}$ is some Euclidean space or its subspace of nonzero Lebesque measure: $\mathbf{A} \subseteq R^m$, $\mathrm{mes}(\mathbf{A}) > 0$. In practice, however, (because of "digitalization" of our real world) the statisticians need to use discrete-valued models of time series, when the observation space $\mathbf{A}$ is some discrete set with cardinality $N = |\mathbf{A}|$, $\mathrm{mes}(\mathbf{A}) = 0$. Give some applied areas where discrete-valued time series models are extremely helpful (Weiss 2018): bioinformatics for analysis of genetic sequences ($N=4$); information systems for information protection ($N = 2$); meteorology for weather prediction; social science for modelling of dynamics in social behavior; public health and personalized medicine; prediction of environmental processes; financial engineering; telecommunications; alarm systems.

Discrete-valued time series is a random process $x_t \in \mathbf{A}$ on some probability space $(\mathbf{\Omega}, \mathbf{F}, \mathbf{P})$ with discrete time $t \in \mathbf{N}_0 = \{0, 1, 2, \ldots\}$ and some discrete state space $\mathbf{A}$ with the cardinality $N = |\mathbf{A}|$, $2 \le N \le +\infty$. Much attention to discrete-valued time series considered as categorical time series was paid by B. Kedem and K. Fokianos (Kedem and Fokianos 2002).

If $\mathbf{A}$ is any countable set ($N = +\infty$), then we have a countable valued time series x_t. If $\mathbf{A}$ is any finite set ($N < +\infty$), then we have a finitely valued time series, also called categorical time series (Kedem and Fokianos 2002). We will consider here these cases both, and, without

loss of generality, we will assume that $\mathbf{A} = \{0, 1, \ldots, N-1\}$.

An universal base model for discrete-valued time series $x_t \in \mathbf{A}$ is the homogeneous Markov chain MC(s) of some order $s \in \mathbf{N}_0$, determined by the generalized Markov property ($t > s$):

$$\mathbf{P}\{x_t{=}i_t|x_{t-1}{=}i_{t-1}, \ldots, x_1{=}i_1\}{=}\mathbf{P}\{x_t{=}i_t|x_{t-1}{=}i_{t-1}, \ldots, x_{t-s}{=}i_{t-s}\}{=}p_{i_{t-s},\ldots,i_{t-1},i_t}, \quad (1)$$

where s is the memory depth; $i_1, i_2, \ldots, i_t \in \mathbf{A}$ are values of the process at the time moments $1, 2, \ldots, t$ respectively; $P = \left(p_{i_{t-s},\ldots,i_{t-1},i_t}\right)$ is an $(s+1)$-dimensional matrix of one-step transition probabilities. Number of independent parameters for the MC(s) model increases exponentially w.r.t. the memory depth s: $D_{\text{MC}(s)} = N^s(N-1)$. To identify this model (1) we need to have huge data set and the computation work of size $O\left(N^{s+1}\right)$. To avoid this "curse of dimensionality" we propose to use the parsimonious ("small-parametric") models of high-order Markov chains that are determined by small number of parameters $d \ll D_{\text{MC}(s)}$ (Kharin 2013).

2. Construction of parsimonious high-order Markov chains

According to what has been said in Introduction, the parsimonious high-order Markov chain is determined by parsimonious representation of the one-step transition probabilities matrix P defined by (1). Number of independent parameters d of the parsimonious matrix is much smaller than the total number of independent parameters and is determined by the *parsimony coefficient*:

$$\varkappa ::= \frac{d}{D_{\text{MC}(s)}} \ll 1. \quad (2)$$

In our opinion, there are two main approaches to construction of parsimonious matrix P:

Approach I: squeezing of the set of different values of elements in matrix P;

Approach II: using of some generation equation for the conditional probability distribution (1) of the future state x_t under its prehistory.

To define **Approach I** let us introduce some notations: $Q = \left(q_{j_1,\ldots,j_r,j_{r+1}}\right)$ is some stochastic $(r+1)$-dimensional matrix, $1 \le r < s$, $\sum_{j_{r+1} \in \mathbf{A}} q_{j_1,\ldots,j_r,j_{r+1}} \equiv 1$, $0 \le q_{j_1,\ldots,j_r,j_{r+1}} \le 1$, $j_1, \ldots, j_{r+1} \in \mathbf{A}$; $\mathbf{B}(\cdot) : \mathbf{A}^s \to \mathbf{A}^r$ is some discrete function. The $(s+1)$-dimensional matrix P is squeezed to the $(r+1)$-dimensional matrix Q by the general transformation:

$$p_{i_1,\ldots,i_s,i_{s+1}} = q_{\mathbf{B}(i_1,\ldots,i_s),i_{s+1}} \quad (3)$$

with the parsimony coefficient (2) equal to

$$\varkappa = N^{r-s} < 1. \quad (4)$$

Illustration of the squeezing approach (3) for the case of $N = 2$, $s = 4$, $r = 2$ is given by Figure 1 and Figure 2. As it is seen from these figures and from (3), matrix P has many identical rows (they are marked by identical colors).

Give a list of known parsimonious high-order Markov chains correspondent to **Approach I**: Markov chain of order s with r partial connections MC(s, r) (Kharin and Piatlitski 2007), Markov chain of conditional order MCCO(s, r) (Kharin and Maltsew 2017), variable length Markov chain (Buhlmann and Wyner 1999).

Approach II to construction of parsimonious high-order Markov chain is based on some generation equation for the conditional probability distribution of the future state $x_t \in \mathbf{A}$ under its prehistory $X_{t-s}^{t-1} = (x_{t-1}, \ldots, x_{t-s})'$:

$$p_{i_1,\ldots,i_s,i_{s+1}} = q_{i_{s+1}}\left(\theta(i_1,\ldots,i_s;a)\right), \ i_1, \ldots, i_{s+1} \in \mathbf{A}, \quad (5)$$

where $\{q_j(\theta) : j \in \mathbf{A}\}$ is some discrete probability distribution on $\mathbf{A}$ that is dependent on the parameter $\theta = (\theta_j) \in \Theta \subseteq R^L$, $\theta(i_1, \ldots, i_s; a)$ is some parametric function that is a priori known up to some unknown vector parameter $a = (a_k) \in R^m$.

$$\mathbf{P} = \begin{array}{cccc} & & & \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array} \begin{pmatrix} 0 & 1 \\ q_{000} & 1-q_{000} \\ q_{000} & 1-q_{000} \\ q_{010} & 1-q_{010} \\ q_{010} & 1-q_{010} \\ q_{000} & 1-q_{000} \\ q_{000} & 1-q_{000} \\ q_{010} & 1-q_{010} \\ q_{010} & 1-q_{010} \\ q_{100} & 1-q_{100} \\ q_{100} & 1-q_{100} \\ q_{110} & 1-q_{110} \\ q_{110} & 1-q_{110} \\ q_{100} & 1-q_{100} \\ q_{100} & 1-q_{100} \\ q_{110} & 1-q_{110} \\ q_{110} & 1-q_{110} \end{pmatrix}$$

Figure 1: Parsimonious matrix P: $N=2$, $s=4$, $r=2$, $\varkappa=0.25$.
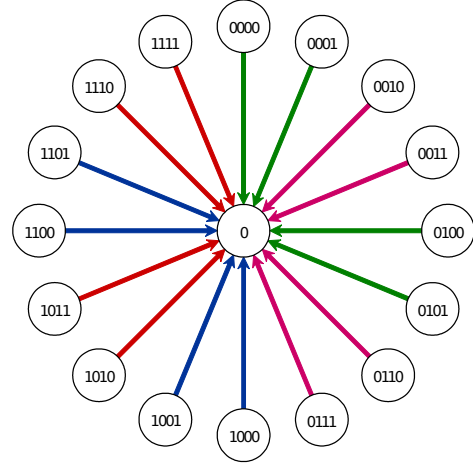
Figure 2: Visualization of the parsimonious matrix from Figure 1.

The parsimony coefficient for this approach is

$$\varkappa = \frac{m}{N^s(N-1)} \leq 1. \tag{6}$$

List of known parsimonious high-order Markov chains correspondent to **Approach II** is longer: Jakobs – Lewis model (Jacobs and Lewis 1978), MTD-model (Raftery 1985), DAR(s) (Kharin and Maltsew 2017), BCNAR(s) (Kharin, Voloshko, and Medved 2018), BiCNAR(s) (Kharin and Voloshko 2019), PCNAR(s) (Kharin and Kislach 2019), INAR(s) (Alzaid and Al-Osch 1990).

3. Statistical analysis for models constructed by Approach I

3.1. Markov chain MC(s, r) of order s with r partial connections

The MC(s, r) proposed by Yu. Kharin in 2004 is determined by the following parsimonious presentation of the $(s+1)$-dimensional transition probability matrix:

$$p_{J_1^{s+1}} = p_{j_1,\ldots,j_s,j_{s+1}} = q_{j_{m_1^0},\ldots,j_{m_r^0},j_{s+1}}, \tag{7}$$

where $J_1^{s+1} = (j_1,\ldots,j_{s+1}) \in \mathbf{A}^s$ is the $(s+1)$-dimensional index vector; r is the number of connections ($1 \leq r \leq s$); $M_r^0 = \left(m_1^0,\ldots,m_r^0\right) \in M$ is some integer-valued vector with r ordered components, $1 = m_1^0 < m_2^0 < \ldots < m_r^0 \leq s$, called the template of connections; $Q = \left(q_{J_1^{r+1}}\right)_{J_1^{r+1} \in \mathbf{A}^{r+1}}$ is an $(r+1)$-dimensional stochastic matrix. If $r = s$, we have the general MC(s)-model.

In (Kharin 2013; Kharin and Piatlitski 2007) the following probabilistic properties of the MC(s, r)-model are found.

Theorem 1. *The MC(s, r) defined by (7) is an ergodic Markov chain iff there exists $i \in N$ such that* $\min\limits_{J_1^s, J_{s+i+1}^{2s+i} \in \mathbf{A}^s} \sum\limits_{J_{s+1}^{s+i} \in \mathbf{A}^i} \prod\limits_{k=1}^{s+i} q_{j_{k+m_1^0-1},\ldots,j_{k+m_r^0-1},j_{k+s}} > 0$. *Stationary probability distribution* $\pi_{J_1^s}^*$ *satisfies the equations:* $\pi_{J_2^{s+1}}^* = \sum\limits_{j_1 \in \mathbf{A}} \pi_{J_1^s}^* q_{j_{m_1^0},\ldots,j_{m_r^0},j_{s+1}}$, $J_1^{s+1} \in \mathbf{A}^s$.

Corollary 1. *For a stationary Markov chain the stationary probability distribution has the multiplicative form:*

$$\pi^*_{J_1^s} = \prod_{i=1}^{s} \pi^*_{j_i}, \quad J_1^s \in \mathbf{A}^s, \quad iff \quad \pi^*_{j_{r+1}} = \sum_{j_1 \in \mathbf{A}} \pi^*_{j_1} q_{J_1^{r+1}}, \; J_2^{r+1} \in \mathbf{A}^r.$$

Introduce the notation: δ_{ij} is Kronecker symbol; $F\left(J_i^{i+s-1}; M_r\right) = (j_{i+m_1-1}, \ldots, j_{i+m_r-1})$ is the selector-function; $X_1^n = (x_1, \ldots, x_m)'$ is the observed time series of length n;

$$\nu_{J_1^{r+1}}\left(X_1^n; M_r\right) = \sum_{t=1}^{n-s} \delta_{F\left(X_t^{t+s-1}; M_r\right), J_1^r} \delta_{x_{t+s}, j_{r+1}} \tag{8}$$

is the frequency statistic for the template $M_r \in M$; $\mu_{J_1^{r+1}}(M_r) = \mathbf{P}\{F\left(X_t^{t+s-1}; M_r\right) = J_1^r, x_{t+s} = j_{r+1}\}$ is the probability distribution of the $(r+1)$-tuple; the dot used instead of any index means summation on all its values: $\mu_{J_1^r \bullet}(M_r) = \sum_{j_{r+1} \in \mathbf{A}} \mu_{J_1^{r+1}}(M_r)$; $\hat{\mu}_{J_1^{r+1}}(M_r) = \nu_{J_1^{r+1}}\left(X_1^n; M_r\right)/(n-s)$ is the frequency estimator for the probability $\mu_{J_1^{r+1}}(M_r)$, $J_1^{r+1} \in \mathbf{A}^{r+1}$, $M_r \in M$.

We have the following results (Kharin and Piatlitski 2007).

Theorem 2. *If the template of connections M_r^0 is known, then the maximum likelihood estimator (MLE) for the matrix Q is $\hat{Q} = \left(\hat{q}_{J_1^{r+1}}\right)_{J_1^{r+1} \in \mathbf{A}^{r+1}}$: $\hat{q}_{J_1^{r+1}} = \left\{\hat{\mu}_{J_1^{r+1}}(M_r^0)/\hat{\mu}_{J_1^r \bullet}(M_r^0)$, if $\hat{\mu}_{J_1^r \bullet}(M_r^0) > 0; 1/N$ else$\right\}$. Under the stationarity condition, $\left\{\hat{q}_{J_1^{r+1}} : J_1^{r+1} \in \mathbf{A}^{r+1}\right\}$ are asymptotically ($n \to \infty$) unbiased and consistent with covariances*

$$\mathrm{Cov}\left\{\hat{q}_{J_1^{r+1}}, \hat{q}_{K_1^{r+1}}\right\} = \sigma^{\hat{q}}_{J_1^{r+1}, K_1^{r+1}}/(n-s) + O(1/n^2),$$

$$\sigma^{\hat{q}}_{J_1^{r+1}, K_1^{r+1}} = \delta_{J_1^r, K_1^r} q_{J_1^{r+1}} \left(\delta_{j_{r+1}, k_{r+1}} - q_{K_1^{r+1}}\right)/\mu_{J_1^r \bullet}(M_r^0), \quad J_1^{r+1}, K_1^{r+1} \in \mathbf{A}^{r+1}.$$

Moreover, the probability distribution of the N^{r+1}-dimensional composed random vector $\left(\sqrt{n-s}\left(\hat{q}_{J_1^{r+1}} - q_{J_1^{r+1}}\right)\right)_{J_1^{r+1} \in \mathbf{A}^{r+1}}$ at $n \to \infty$ converges to the normal probability distribution with zero mean and the covariance matrix $\Sigma^{\hat{q}} = \left(\sigma^{\hat{q}}_{J_1^{r+1}, K_1^{r+1}}\right)$.

The consistent statistical test for hypotheses: $H_0 = \{Q = Q^0\}$, where $Q^0 = \left(q^0_{J_1^{r+1}}\right)_{J_1^{r+1} \in \mathbf{A}^{r+1}}$ is some given matrix; $H_1 = \overline{H}_0$, consists of the following steps.

1. Computation of the statistics $\nu_{J_1^{r+1}}(X_1^n; M_r^0)$, $J_1^{r+1} \in \mathbf{A}^{r+1}$, by (8).

2. Computation of the statistic $\left(D_{J_1^r} = \left\{j_{r+1} \in \mathbf{A} : q^0_{J_1^{r+1}} > 0\right\}\right)$

$$\rho = \sum_{J_1^r \in \mathbf{A}^r, \, j_{r+1} \in D_{J_1^r}} \nu_{J_1^r \bullet}(X_1^n; M_r^0) \left(\hat{q}_{J_1^{r+1}} - q^0_{J_1^{r+1}}\right)^2 / q^0_{J_1^{r+1}}.$$

3. Computation of the P-value: $P = 1 - G_U(\rho)$, where $G_U(\cdot)$ is the standard χ^2- cumulative distribution function with $U = \sum_{J_1^r \in \mathbf{A}^r} \left(|D_{J_1^r}| - 1\right)$ degrees of freedom.

4. The decision rule with an asymptotic significance level ε: if $P \geq \varepsilon$, then to conclude that the hypothesis H_0 is true; otherwise, the alternative H_1 is true.

Corollary 2. *Under stationary $MC(s,r)$ and contigual family of alternatives $H_{1n} = \{Q =$*
$Q^{1n}\}$, $q_{J_1^{r+1}}^{1n} = \left(q_{J_1^{r+1}}^{1n}\right)_{J_1^{r+1} \in \mathbf{A}^{r+1}}$, $q_{J_1^{r+1}}^{1n} = q_{J_1^{r+1}}^0 \left(1 + \dfrac{d_{J_1^{r+1}}}{\sqrt{n-s}}\right)$, $\displaystyle\sum_{j_{r+1} \in \mathbf{A}} d_{J_1^{r+1}} q_{J_1^{r+1}}^0 = 0$,
$\displaystyle\sum_{J_1^{r+1} \in \mathbf{A}^{r+1}} \left|d_{J_1^{r+1}}\right| > 0$, *if H_{1n} is true, then at $n \to \infty$ the power of the developed test*
$w \to 1 - G_{U,a}\left(G_U^{-1}(1-\varepsilon)\right)$, *where $G_{U,a}(\cdot)$ is the cumulative distribution function of the*
noncentral χ^2-distribution with U degrees of freedom and the noncentrality parameter

$$a = \sum_{J_1^{r+1} \in \mathbf{A}^{r+1}} \mu_{J_1^{r+1}}(M_r^0) d_{J_1^{r+1}}^2.$$

Introduce the notation: M is the set of all admissible templates M_r;

$$H(M_r) = - \sum_{J_1^{r+1} \in \mathbf{A}^{r+1}} \mu_{J_1^{r+1}}(M_r) \ln\left(\mu_{J_1^{r+1}}(M_r)/\mu_{J_1^{r+1} \bullet}(M_r)\right) \geq 0 \qquad (9)$$

is the conditional entropy of the future symbol $x_{t+s} \in \mathbf{A}$ relative to the past derived by the
selector $F\left(X_t^{t+s-1}; M_r\right) \in \mathbf{A}^r$, $M_r \in M$; $\hat{H}(M_r)$ is the "plug-in" estimator of the condi-
tional entropy, which is generated by substitution of the estimators $\hat{\mu}_{J_1^{r+1}}(M_r)$ instead of true
probabilities $\mu_{J_1^{r+1}}(M_r)$ into (9).

Theorem 3. *If the order s and the number of connections r are known, then the MLE*
$\hat{M}_r = \arg\min_{M_r \in M} \hat{H}(M_r)$. *Under the stationarity condition of the $MC(s,r)$ the estimator $\hat{M}_r$*
at $n \to \infty$ is consistent: $\hat{M}_r \xrightarrow{\mathbf{P}} M_r^0$.

Let $s \in [s_-, s_+]$, $r \in [r_-, r_+]$, $1 \leq s_- < s_+ < \infty$, $1 \leq r_- < r_+ < s_+$. To estimate parameters
r, s we use the Bayesian Information Criterion (BIC):

$$BIC(s,r) = 2(n-s)\hat{H}(\hat{M}_r) + U\ln(n-s), \qquad (10)$$

where $U = \displaystyle\sum_{J_1^r \in \mathbf{A}^r} \left(\left|D_{J_1^r}\right| - 1 + \delta_{\hat{\mu}_{J_1^r \bullet}(\hat{M}_r), 0}\right)$, $D_{J_1^r} = \left\{j_{r+1} \in \mathbf{A} : \hat{\mu}_{J_1^{r+1}}(\hat{M}_r) > 0\right\}$.

Consistent estimators $\hat{s}$, $\hat{r}$ are determined by minimization: $\mathrm{BIC}(s,r) \to \min\limits_{s,r}$.

3.2. Markov chain of conditional order MCCO(s, L)

Introduce the notation: $L \in \{1, \dots, s-1\}$ is some positive integer, $K = N^L - 1$; $Q^{(1)}, \dots, Q^{(M)}$
are M $(1 \leq M \leq K+1)$ different square stochastic matrices of the order N:

$$Q^{(m)} = \left(q_{i,j}^{(m)}\right), \; 0 \leq q_{i,j}^{(m)} \leq 1, \; \sum_{j \in \mathbf{A}} q_{i,j}^{(m)} \equiv 1, \; i, j \in \mathbf{A}, \; 1 \leq m \leq M;$$

$< J_n^m > = \displaystyle\sum_{k=n}^m N^{k-n} j_k \in \{0, 1, \dots, N^{m-n+1} - 1\}$ is the numeric representation of the multiin-
dex $J_n^m \in \mathbf{A}^{m-n+1}$; $\mathrm{I}\{C\}$ is the indicator function of the event C; $1 \leq m_k \leq M$, $1 \leq b_k \leq s-L$,
$0 \leq k \leq K$. It is assumed that sequences $\{m_k\}$ and $\{b_k\}$ are fixed, $\min\limits_{0 \leq k \leq K} b_k = 1$ and all ele-
ments of the set $\{1, 2, \dots, M\}$ are present in the sequence $m_0, \dots, m_K$.

Markov chain $\{x_t \in \mathbf{A} : t \in \mathbb{N}\}$ is called (Kharin and Maltsew 2017) the Markov chain of
conditional order (MCCO(s, L)), if its one-step transition probabilities have the following
parsimonious form:

$$p_{J_1^{s+1}} = \sum_{k=0}^K \mathrm{I}\{< J_{s-L+1}^s >= k\} q_{j_{b_k}, j_{s+1}}^{(m_k)}.$$

The sequence of elements J_{s-L+1}^s is called the base memory fragment (BMF) of the random sequence; L is the length of BMF; the value $s_k = s - b_k + 1$ is called the conditional order of Markov chain. Thus the conditional probability distribution of the state x_{t+1} at time point $t+1$ depends not on all s previous states, but it depends only on $L+1$ selected states $\left(j_{b_k}, J_{s-L+1}^s\right)$. Note that if $L = s - 1$, $s_0 = s_1 = \ldots = s_K = s$, we have the full-connected Markov chain of the order s: $MC(s)$. If $M = K + 1$, then each transition matrix corresponds to only one value of the BMF, otherwise there exists a common matrix which corresponds to several values of BMF.

Hence the transition matrix P of the Markov chain of conditional order is determined by $d = 2(N^L + 1) + MN(N - 1)$ independent elements, and the parsimony coefficient (2) is

$$\varkappa = \frac{N^L + MN(N-1)}{N^s(N-1)} \leq 1.$$

Methods and algorithms of statistical analysis for $MCCO(s, L)$ are presented in (Kharin and Maltsew 2017).

4. Statistical analysis for models constructed by Approach II

4.1. Jacobs - Lewis model

Jacobs – Lewis model is determined by a stochastic difference equation (Jacobs and Lewis 1978) $(t > s)$:

$$x_t = \mu_t x_{t-\eta_t} + (1 - \mu_t)\xi_t, \tag{11}$$

where $\{\xi_t, \eta_t, \mu_t\}$ are independent random variables with probability distributions:

$$\mathbf{P}\{\mu_t = 1\} = 1 - \mathbf{P}\{\mu_t = 0\} = \rho; \quad \mathbf{P}\{\xi_t = k\} = \pi_k, \quad k \in \mathbf{A}, \quad \sum_{k \in \mathbf{A}} \pi_k = 1;$$

$$\mathbf{P}\{\eta_t = i\} = \lambda_i, \quad i \in \{1, 2, \ldots, s\}, \quad \sum_{i=1}^s \lambda_i = 1, \quad \lambda_s \neq 0. \tag{12}$$

Number of parameters depends linearly on s: $D_{\mathrm{JL}} = N + s - 1$; $\varkappa = (N + s - 1)/(N^s(N - 1)) \leq 1$.

In (Jacobs and Lewis 1978) only moments and stationary distributions were analyzed. We proved (Kharin 2013) probabilistic and statistical properties of the model (11), (12) by the $MC(s)$-model.

Theorem 4. *Discrete-valued time series x_t determined by (11), (12) is a homogeneous Markov chain of the order s with the initial probability distribution $\pi_{i_1,\ldots,i_s} = \pi_{i_1} \cdot \ldots \cdot \pi_{i_s}$ and the $(s+1)$-dimensional matrix of transition probabilities $P(\pi, \lambda, \rho) = \left(p_{i_1,\ldots,i_{s+1}}\right)$:*

$$p_{i_1,\ldots,i_s,i_{s+1}} = (1 - \rho)\pi_{i_{s+1}} + \sum_{j=1}^s \lambda_j \delta_{i_{s-j+1}, i_{s+1}}, \quad i_1, \ldots, i_{s+1} \in \mathbf{A}.$$

Corollary 3. *Maximum likelihood estimators $\hat{\pi}, \hat{\lambda}, \hat{\rho}$ by the data $X_1^n = (x_1, \ldots, x_n)'$ are determined by the solution of the maximization problem:*

$$l(\pi, \lambda, \rho) = \sum_{t=1}^s \ln \pi_{x_t} + \sum_{t=s+1}^n \ln \left((1 - \rho)\pi_{x_t} + \rho \sum_{j=1}^s \lambda_j \delta_{x_{t-j}, x_t}\right) \to \max_{\pi, \lambda, \rho}.$$

Using these MLEs we had constructed (Kharin 2013) the consistent generalized probability ratio test for hypotheses on the true values of parameters ρ, λ, π in (12) as in Subsection 3.1.

4.2. Raftery model

Mixture Transition Distribution-model (MTD-model) was proposed by A. Raftery (Raftery 1985) as a special parsimonious representation of the matrix P:

$$p_{i_1,\ldots,i_s,i_{s+1}} = \sum_{j=1}^{s} \lambda_j q_{i_j,i_{s+1}}, \quad i_1,\ldots,i_{s+1} \in \mathbf{A}, \tag{13}$$

where $Q = (q_{i,k})$ is a stochastic $(N \times N)$-matrix, $0 \le q_{i,k} \le 1$, $\sum_{k\in\mathbf{A}} q_{i,k} \equiv 1$, $i,k \in \mathbf{A}$, $\lambda = (\lambda_1,\ldots,\lambda_s)'$ is a discrete probability distribution, $\lambda_1 > 0$.

The MTDg (generalized MTD)-model is:

$$p_{i_1,\ldots,i_s,i_{s+1}} = \sum_{j=1}^{s} \lambda_j q^{(j)}_{i_{s-j+1},i_{s+1}}, \quad i_1,\ldots,i_{s+1} \in \mathbf{A}, \tag{14}$$

where $Q^{(j)} = \left(q^{(j)}_{i,k}\right)$ is a stochastic matrix for the j-th lag.

Number of parameters for the MTDg: $D_{\mathrm{MTDg}} = s(N(N-1)/2+1)-1$; $\varkappa = O(s \cdot N^{2-s})$.

We have constructed a simple criterion for the ergodicity of the MTD-model and found a useful property of the stationary probability distribution (Kharin 2013).

Theorem 5. *For the MTDg-model (14), if $\exists K \in N : \left(\left(Q^{(1)}\right)^K\right)_{ij} > 0, \forall i,j \in \mathbf{A}$, then the s-dimensional stationary probability distribution satisfies the equation $(i_1,\ldots,i_s \in \mathbf{A})$:*

$$\pi^*_{i_1,\ldots,i_s} = \prod_{l=0}^{s-1} \left(\pi^*_{i_s-l} + \sum_{j=l+1}^{s} \lambda_j \left(q^{(j)}_{i_{j-l},i_{s-l}} - \sum_{r=0}^{N-1} q^{(j)}_{r,i_{s-l}} \pi^*_r \right) \right).$$

Corollary 4. *For the ergodic MTD-model (13) the 2-dimensional stationary probability distribution of the random vector $(x_{t-m},x_t)'$ is $\pi^*_{ki}(m) = \pi^*_k\pi^*_i + \pi^*_k\lambda_{s-m+1}(q_{ki}-\pi^*_i)$, $i,k \in \mathbf{A}$, $1 \le m \le s$.*

Based on Corollary 4 we construct statistical estimators $\tilde{\lambda}$, $\tilde{Q}$ by an observed time series $X_1^n = (x_1,\ldots,x_n)'$ of the length n:

$$\tilde{\pi}_i = \frac{1}{n-2s+1} \sum_{t=s+1}^{n-s+1} \delta_{x_t,i}; \quad \tilde{\pi}_{ki}(j) = \frac{1}{n-2s+1} \sum_{t=s+j}^{n-s+j} \delta_{x_{t-j},k}\delta_{x_t,i};$$

$$z_{ki}(j)=\tilde{\pi}_{ki}(s-j)/\tilde{\pi}_k-\tilde{\pi}_i, \quad d_{ki}=\tilde{q}_{ki}-\tilde{\pi}_i, \quad i,k\in\mathbf{A}; \quad \tilde{\lambda}_j = \sum_{i,k\in\mathbf{A}} z_{ki}(s-j)d_{ki} / \sum_{i,k\in\mathbf{A}} d^2_{ki}, \quad j=1,\ldots,s;$$

$$\tilde{q}_{ki} = \left\{ \sum_{j=1}^{s} \tilde{\pi}_{ki}(j)/\tilde{\pi}_k - (s-1)\tilde{\pi}_i, \quad \text{if} \quad \tilde{\pi}_k > 0; \quad N^{-1} \quad \text{else} \right\}. \tag{15}$$

Theorem 6. *For the ergodic MTD-model (13) the estimators $\tilde{Q}$, $\tilde{\lambda}$ determined by (15) at $n \to \infty$ are consistent and asymptotically unbiased.*

MLE $\hat{Q}$, $\hat{\lambda}$ are solutions of the nonlinear maximization problem:

$$l(Q,\lambda) = \sum_{t=s+1}^{n} \ln \sum_{j=1}^{s} \lambda_j q_{x_{t-s+j-1},x_t} \to \max_{Q,\lambda}. \tag{16}$$

The estimators $\tilde{Q}$, $\tilde{\lambda}$ are used as initial values in the iterative computation of the MLEs $\hat{Q}$, $\hat{\lambda}$ in (16). Generalized probability ratio test of the asymptotic size $\varepsilon \in (0,1)$ for $H_0 = \{Q = Q^0, \lambda = \lambda^0\}$, $H_1 = \overline{H}_0$ is constructed as in the Subsection 3.1.

4.3. Binomial conditionally nonlinear autoregressive model BiCNAR(s)

This model (Kharin and Voloshko 2019) is determined by the special (binomial) case of equation (5):

$$p_{i_1,\ldots,i_s,i_{s+1}} = C_{N-1}^{i_{s+1}} \theta^{i_{s+1}} (1-\theta)^{N-1-i_{s+1}}, \ i_{s+1} \in \mathbf{A} = \{0,1,\ldots,N-1\}, \qquad (17)$$

$$\theta = \theta\left(\mathrm{I}_1^s\right) = \mathrm{F}\left(a'\Psi\left(\mathrm{I}_1^s\right)\right), \ \mathrm{I}_1^s = (i_1,\ldots,i_s)' \in \mathbf{A}^s.$$

Here $\Psi\left(\mathrm{I}_1^s\right) = \left(\psi_1\left(\mathrm{I}_1^s\right),\ldots,\psi_m\left(\mathrm{I}_1^s\right)\right)' : \mathbf{A}^s \to R^m$ is a column-vector of $m \le N^s$ linearly independent functions, e.g. polynomials, $\mathrm{F}(\cdot): R^1 \to [0,1]$ is some fixed cumulative distribution function, e.g. logistic, normal or Cauchy:

$$\Lambda(\zeta) = \frac{1}{1+e^{-\zeta}}, \ \Phi(\zeta) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{\zeta} e^{-\frac{x^2}{2}} \, dx, \ C(\zeta) = \frac{1}{2} + \frac{\arctan(\zeta)}{\pi}, \ \zeta \in R^1,$$

$a = (a_1,\ldots,a_m)'$ is a column-vector of m unknown model parameters. Parsimony coefficient for this model $\varkappa = m\left(N^s(N-1)\right)^{-1} \le 1$.

Introduce the notation: $\mathrm{F}^{-1}(\cdot)$ is the quantile function; $X_1^T = (x_1,\ldots,x_T)' \in \mathbf{A}^T$ is the observed time series of length T;

$$\hat{\theta}(J) = \frac{1}{N-1} \cdot \frac{\sum\limits_{t=s+1}^{T} x_t \mathrm{I}\{X_{t-s}^{t-1} = J\}}{\sum\limits_{t=s+1}^{T} \mathrm{I}\{X_{t-s}^{t-1} = J\}}, \ J \in \mathbf{A}^s;$$

$B = (b_J)$ is $(N^s \times 1)$ – vector-column, $b_J = \mathrm{F}^{-1}(\hat{\theta}(J))$; $H = (h_{J,J'})$ is some fixed $(N^s \times N^s)$ – symmetrical non-negatively defined matrix; $\boldsymbol{\Psi} = (\Psi(J))$ is $(m \times N^s)$ matrix; O_m is the zero m-vector.

Theorem 7. *If* $\mathrm{F}(\cdot)$ *satisfies the smoothness assumptions:* $0<\mathrm{F}(\zeta)<1$, $0<\mathrm{F}'(\zeta)< +\infty$, $\mathrm{F}(\cdot)$ *and* $\mathrm{F}^{-1}(\cdot)$ *are twice differentiable, and* $|\boldsymbol{\Psi}H\boldsymbol{\Psi}'|\neq0$, *then the Frequencies Based Estimator (FBE)*

$$\hat{a} = (\boldsymbol{\Psi}H\boldsymbol{\Psi}')^{-1}\boldsymbol{\Psi}HB \qquad (18)$$

is consistent and asymptotically normal at $T \to +\infty$:

$$\hat{a} \xrightarrow{\mathbf{P}} a, \ \sqrt{T}(\hat{a} - a) \xrightarrow{\mathbf{D}} \mathcal{N}_m(O_m, \Sigma_H),$$

$$\Sigma_H = (\boldsymbol{\Psi}H\boldsymbol{\Psi}')^{-1}\boldsymbol{\Psi}HJ^{-1}H(\boldsymbol{\Psi}H\boldsymbol{\Psi}')^{-1},$$

where J *is the Fisher information matrix.*

Corollary 5. *For the weight matrix* $H = J$ *the FBE (18) is asymptotically efficient (it attains the Cramer – Rao boundary at* $T \to +\infty$).

Proofs of Theorem 7 and its corollary are given in (Kharin and Voloshko 2019).

FBE (17) has the following significant advantages w.r.t. the MLE: 1) explicit expression of FBE w.r.t. the iterative computation of MLE; 2) fast iterative computation of FBE if we extend the basis $\{\psi_i(\cdot)\}$; 3) possibility to control the computational complexity of FBE by variation of the matrix H.

Note in conclusion, that the developed method of FBE construction for the BiCNAR is successively used for some other models of discrete-valued time series (Kharin and Kislach 2019): semibinomial, Poisson, geometric, negative binomial conditionally nonlinear autoregressive time series.

4.4. Binomial conditionally nonlinear model for spatio-temporal data

We extend the BiCNAR model determined in the previous subsection for the case of spatio-temporal data (Dauhaliova and Kharin 2018). Introduce the notation: $s \in S = \{1, 2, \ldots, n\}$ is the index variable that determines the geographic region (site); n is the number of geographical regions to be analyzed; $x_{g,t} \in \mathbf{A}$ is the discrete random observation at time moment $t \in \mathbf{N}_0$ at the site $g \in G$; $X_t = (x_{1,t}, \ldots, x_{n,t})' \in \mathbf{A}^n$ is the column-vector of all n observations for all sites at time moment t; $Z_{g,t} = (z_{1,g,t}, \ldots, z_{m,g,t})' \in R^m$ is the column-vector of m exogenous variables at time moment t for the site $g \in G$. It is assumed that under fixed prehistory $\{X_1, \ldots, X_{t-1}\}$: a) the random variables $x_{t,1}, \ldots, x_{t,n}$ are conditionally independent; b) conditional probability distribution of $x_{g,t}$ is binomial:

$$\mathbf{P}\{x_{g,t} = l | X_{t-1}, \ldots, X_1\} = C_{N-1}^l \cdot p_{g,t}^l \cdot (1 - p_{g,t})^{N-1-l}, \quad l \in \mathbf{A}; \tag{19}$$

c) parameter $p_{s,t}$ is determined by special case of generation equation (5):

$$\ln \frac{p_{g,t}}{1 - p_{g,t}} = \sum_{k=1}^{s} \sum_{i=1}^{n} a_{gki} x_{i,t-k} + \sum_{j=1}^{m} b_{g,j} z_{j,g,t}, \quad g \in G, \quad t \in \mathbf{N}_0, \tag{20}$$

where $a_g = \{a_{gki}\} \subset R^m$, $b_g = (b_{g,1}, \ldots, b_{s,m}) \in R^m$ are parameters of this model.

The proposed model (19), (20) determines the parsimonious nonhomogeneous n-dimensional Markov chain of order s with parsimony coefficient:

$$\varkappa = \frac{n(sN + m)}{N^{ns}(N^n - 1)}.$$

We constructed the consistent and asymptotically normal MLEs for parameters $\{a_g\}$, $\{b_g\}$ of the model (19), (20) and numerical procedure for its evaluation in (Kharin and Zhurak 2015; Dauhaliova and Kharin 2018).

Note in conclusion, that in (Kharin and Zhurak 2015) we have developed the Poisson conditionally nonlinear model for spatio-temporal count data.

5. Application of the developed algorithms in data analysis

All algorithms developed in Sections 2-4 were tested on simulated data and illustrated their accordance with the proved theoretical properties. Here we give some results of computer experiments with the developed algorithms on real data from applications. We present here three examples.

5.1. Modeling of wind direction

The discrete-valued time series of the daily average wind speed at Malin Head (North of Ireland) during the period 1961 – 1978 (Jacobs and Lewis 1978) $x_t \in \{0, 1, 2\}$, $N = 3$, of the length $T = 6574$ was fitted by the MC(s, r)-model for $s = \{1, 2, \ldots, 7\}$, $r = \{1, 2, \ldots, 7\}$. Table 1 presents the values of the BIC for different pairs (s, r).

The best fitted model is the MC(3,2) with $\hat{M}_r = (1, 3)$ and the transposed matrix

$$\hat{Q}' = \begin{pmatrix} 0.27 & 0.08 & 0 & 0.22 & 0.04 & 0 & 0.21 & 0.02 & 0 \\ 0.73 & 0.86 & 0.63 & 0.78 & 0.82 & 0.52 & 0.79 & 0.72 & 0.43 \\ 0 & 0.06 & 0.37 & 0 & 0.14 & 0.48 & 0 & 0.26 & 0.57 \end{pmatrix}.$$

The fitted model MC(3, 2) detects significant stochastic dependencies in this data.

Table 1: Modelling of the wind speed data

Model	BIC	Model	BIC	Model	BIC	Model	BIC
MC(1,1)	8127.52	MC(4,2)	8139.12	MC(5,5)	8621.97	MC(7,1)	9041.43
MC(2,1)	8777.63	MC(4,3)	8164.79	MC(6,1)	9016.23	MC(7,2)	8163.07
MC(2,2)	8096.08	MC(4,4)	8332.77	MC(6,2)	8148.48	MC(7,3)	8197.91
MC(3,1)	8849.90	MC(5,1)	8984.10	MC(6,3)	8190.78	MC(7,4)	8323.19
MC(3,2)	**8079.81**	MC(5,2)	8129.83	MC(6,4)	8350.82	MC(7,5)	8599.09
MC(3,3)	8143.13	MC(5,3)	8177.92	MC(6,5)	8576.92	MC(7,6)	8973.15
MC(4,1)	8956.11	MC(5,4)	8349.62	MC(6,6)	8969.54	MC(7,7)	9575.64

5.2. Analysis of CG-patterns in genome

We took the complete Panthera tigris mitochondrion genome of the length $T = 16990$ (available from NCBI Nucleotide data base, ID EF551003.1) and extracted the binary sequence x_t of its CG-indicators: $x_t = 1$ iff the t'th nucleotide is Guanine or Cytosine, $t = 1, \ldots, T$. Portion of "1" in X_1^T is known as CG-content and plays important role in bioinformatics (Kharin *et al.* 2018).

In order to evaluate individual and pairwise impact of the lagged variables X_{t-s}^{t-1} on x_t we fitted the BCNAR(s)-model (up to $s = 15$) for $N = 2$ with the bilinear bases $\{\psi_i(\cdot)\}$ and the Gaussian c.d.f. $\Phi(\cdot)$.

Two fitted BCNAR-models for $s = 10; 15$ respectively, are ($\zeta_t = (-1)^{x_t}$):

$$\mathbf{P}\{x_t | x_{t-1}, \ldots\} = \Phi(-0.3962 + 0.0313\zeta_{t-1} + 0.0241\zeta_{t-3} + 0.033\zeta_{t-10} +$$

$$+ 0.045\zeta_{t-3}\zeta_{t-6} - 0.0576\zeta_{t-3}\zeta_{t-10}),$$

$$\mathbf{P}\{x_t | x_{t-1}, \ldots\} = \Phi(-0.1319 + 0.022\zeta_{t-1} + 0.0269\zeta_{t-6} + 0.0248\zeta_{t-15} -$$

$$- 0.0434\zeta_{t-6}\zeta_{t-15}),$$

(21)

that are adequate according to BIC. It is seen that in the two fitted models (21) of high orders $s = 10$ and $s = 15$ the future state x_t is significantly influenced by the past states x_{t-i} lagged by multiples of three ($i = 3j$), and even the closest predecessor x_{t-1} has a weaker impact on x_t. Note also that the pairs of the lagged variables in these two models are more influential than the single lagged variables, i.e. the bilinear basis catched some essential high-order statistical dependencies invisible to the linear one. These identified properties of the bilinear models confirm the known in genetics effect of statistical 3-periodicity of protein-coding fragments (Kharin *et al.* 2018); this fact indicates the usefulness of the developed theory for statistical analysis of genetic sequences.

5.3. Dynamics analysis for the incidence rate of children leukemia

Real medical data describes the incidence rate of children leukemia in $n = 3$ cites of Belarus for $T = 25$ years after Chernobyl (normalization of the data was made by the number of residents at the group of risk). We used the Poisson conditionally nonlinear autoregressive spatio-temporal model (Kharin and Zhurak 2015).

Statistical estimators of the parameters obtained by $T = 25$ observations:

$$\hat{\theta}_1 = (-0.06, 0.15)', \ \hat{\theta}_2 = (-0.01, -0.02, 1.51)', \ \hat{\theta}_3 = (0.07, 0.05, 0.82)', \quad (22)$$

determine the models for dynamics of incidence rate, e.g. for the site number $s = 3$ the estimator of expected incidence rate (parameter of the Poisson distribution) is

$$\hat{\lambda}_{3,t} = \exp(0.07 x_{3,t-1} \cdot \mathrm{I}\{t > 1\} + 0.05 x_{2,t} + 0.82). \quad (23)$$

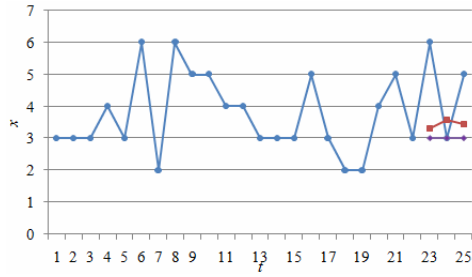Forecasting results by the fitted models are presented in Figures 3-5.
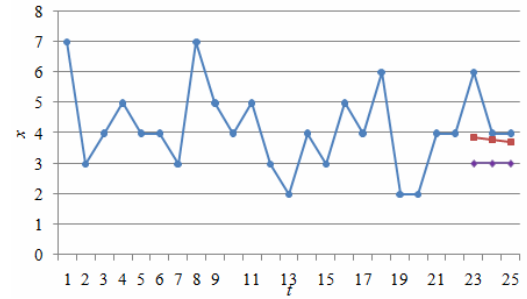
Figure 3: Forecasting results for the site $s=1$.
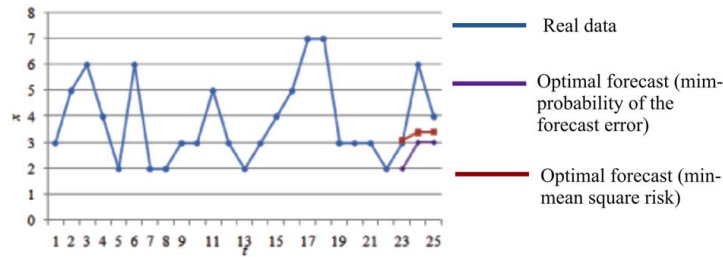

Figure 4: Forecasting results for the site $s=2$.

Figure 5: Forecasting results for the site $s=3$.

6. Conclusion

Methods and software for time series analysis is deep developed for "continuous" data, and these techniques are not (or risky) applicable to discrete-valued time series $x_t \in A$ with discrete state space A.

An universal model for discrete-valued time series with bounded memory length $s \in N$ is the Markov chain of sufficiently large order s, but computational complexity of its fitting by data increases exponentially w.r.t. s: $O(|A|^{s+1})$.

To avoid this "curse of dimensionality" we need to construct parsimonious (small-parametric models). We propose two approaches to develop parsimonious models of high-order Markov chains: 1) squeezing of the set of different values for transition probabilities; 2) using of generation equation for transition probabilities.

The proposed approaches are considered for six parsimonious models.

Developed algorithms of computer data analysis are illustrated in three applications on real data.

Finally, we plan to develop the results of this paper in three main directions: 1) robust statistical analysis of discrete-valued time series based on asymptotical risk expansion methods published in (Kharin 1997, 2011; Kharin and Zhuk 1998); 2) robust sequential testing of hypotheses for discrete-valued time series using approaches from (Kharin 2017; Kharin and Kishylau 2015; Kharin and Tu 2017); 3) applications in steganography (Kharin and Vecherko 2013) and in marketing (Pashkevich and Kharin 2004).

References

Alzaid A, Al-Osch M (1990). "An Integer-valued pth-order Autoregressive Structure (INAR(p)) Process." *Journal of Applied Probability*, **27**, 314–324.

Anderson T (1971). *The Statistical Analysis of Time Series*. Wiley.

Buhlmann P, Wyner A (1999). "Variable Length Markov Chains." *The Annals of Statistics*, **27**(2), 480–513.

Dauhaliova M, Kharin Y (2018). "Asymptotic Analysis of Statistical Estimators of Parameters for Binomial Conditionally Autoregressive Model of Spatio-temporal Data." *Journal of the Belarusian State University. Mathematics and Informatics*, **2**, 47–57.

Jacobs P, Lewis P (1978). "Discrete Time Series Generated by Mixtures I: Correlational and Runs Properties." *Journal of the Royal Statistical Society. Ser. B*, **40**(1), 94–105.

Kedem B, Fokianos K (2002). *Regression Models for Time Series Analysis*. Wiley.

Kharin A (2017). "An Approach to Asymptotic Robustness Analysis of Sequential Tests for Composite Parametric Hypotheses." *Journal of Mathematical Sciences (United States)*, **227**(2), 196–203.

Kharin A, Kishylau D (2015). "Robust Sequential Test for Hypotheses about Discrete Distributions in the Presence of "Outliers"." *Journal of Mathematical Sciences (United States)*, **205**(1), 68–73.

Kharin A, Tu T (2017). "Performance and Robustness Analysis of Sequential Hypotheses Testing for Time Series with Trend." *Austrian Journal of Statistics*, **46**(3-4), 23–36.

Kharin Y (1997). "Robustness of Clustering under Outliers." *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, **1280**, 501–511.

Kharin Y (2011). "Robustness of the Mean Square Risk in Forecasting of Regression Time Series." *Communications in Statistics – Theory and Methods*, **40**(16), 2893–2906.

Kharin Y (2013). *Robustness in Statistical Forecasting*. Springer.

Kharin Y, Kislach M (2019). "Statistical Analysis of Poisson Conditionally Nonlinear Autoregressive Time Series by Frequencies-based Estimators." *Proceedings of the 14-th International Conference. Minsk, Belarus "Pattern Recognition and Information Processing"*, pp. 233–236.

Kharin Y, Maltsew M (2017). "Statistical Analysis of High-order Dependencies." *Acta et Commentationes Universitetas Tartuensis de Mathematica*, **21**(1), 37–45.

Kharin Y, Piatlitski A (2007). "A Markov Chain of Order s with r Partial Connections and Statistical Inference on Its Parameters." *Discrete Mathematics and Applications*, **17**(3), 295–317.

Kharin Y, Vecherko E (2013). "Statistical Estimation of Parameters for Binary Markov Chain Models with Embeddings." *Discrete Mathematics and Applications*, **23**(2), 153–169.

Kharin Y, Voloshko V (2019). "Binomial Conditionally Nonlinear Autoregressive Model of Discrete-valued Time Series and Its Probabilistic and Statistical Properties." *Transactions of the Institute of Mathematics of NAS Belarus*, **26**(1), 95–105.

Kharin Y, Voloshko V, Medved E (2018). "Statistical Estimation of Parameters for Binary Conditionally Nonlinear Autoregressive Time Series." *Mathematical Methods of Statistics*, **26**(2), 103–118.

Kharin Y, Zhuk E (1998). "Filtering of Multivariate Samples Containing "Outliers" for Clustering." *Pattern Recognition Letters*, **19**(12), 1077–1085.

Kharin Y, Zhurak M (2015). "Statistical Analysis of Spatio-temporal Data Based on Poisson Conditional Autoregressive Model." *INFORMATICA*, **26**(1), 67–87.

Pashkevich M, Kharin Y (2004). "Robust Estimation and Forecasting for Beta-Mixed Hierarchical Models of Grouped Binary Data." *SORT*, **28**(2), 125–159.

Raftery A (1985). "A Model for High-order Markov Chains." *Journal of the Royal Statistical Society. Ser. B*, **47**(3), 528–539.

Weiss C (2018). *An Introduction to Discrete-Valued Time Series*. Wiley.

Affiliation:

Yuriy Kharin
Research Institute for Applied Problems of Mathematics and Informatics,
Belarusian State University
220030 Minsk, Belarus
E-mail: Kharin@bsu.by
URL: https://www.bsu.by/ru/main.aspx?guid=3351